Nonparametric Bayesian Kernel Models

FENG LIANG, KAI MAO, MING LIAO, SAYAN MUKHERJEE, and MIKE WEST

Presented by Yingjian Wang
Dec 16. 2011
The regression problem

The general regression model:

\[ y = f(x) + \text{error}. \]
The regularization method

\[ \frac{1}{p} \sum_{i=1}^{p} V(f(x_i), y_i) + \lambda \| f \|_{\mathcal{H}}^2, \quad \lambda > 0 \]
Hilbert space

Complete inner product space:

\[ ||x|| = \sqrt{\langle x, x \rangle} \]

\[ ||x|| > 0 \text{ if } x \neq 0; \]

\[ ||c \cdot x|| = |c| ||x||; \]

\[ ||x+y|| \leq ||x|| + ||y||. \]

Examples of Hilbert space: \( L^2(a, b) \)

\[ \langle f, g \rangle = \int_a^b f(t)g(t)dt \]

\[ ||f|| = \sqrt{\int_a^b |f(t)|^2 dt} \]

What is the angle between \( \sin(t) \) and \( \cos(t) \): \( \pi / 2 \)
Kernel Reproducing Hilbert Space (KRHS)

• The reproducing property:

If $\mathcal{H}$ is a RKHS, then for each $t \in X$ there exists, by the Riesz representation theorem a function $K_t$ of $\mathcal{H}$ (called representer) with the reproducing property

$$\mathcal{F}_t[f] = \langle K_t, f \rangle_{\mathcal{H}} = f(t).$$

• The representer theorem (Kimeldorf and Wahba, 1971):

$$\hat{f}(x) = \sum_{i=1}^{n} w_i k(x, x_i) \quad n \ll p$$
Nonparametric Bayesian framework

\[ \mathcal{H}_k = \left\{ f \mid f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x) \text{ such that } \sum_{j=1}^{\infty} \frac{a_j^2}{\lambda_j} < \infty \right\} \]

\[ \lambda_j \phi_j(x) = \int_{\mathcal{X}} k(x, u) \phi_j(u) \, d\mu(u) \]

\[ \mathcal{G} = \left\{ f \mid f(x) = \int k(x, u) \, d\gamma(u), \ \gamma \in \Gamma \right\} \]

Equivalences between \( \mathcal{G} \) and \( \mathcal{H}_k \) exist for Dirichlet process priors.
Dirichlet process prior

Change the expression a little bit:

\[
f(x) = \int k(x, u) \, d\gamma(u) = \int k(x, u) \, w(u) \, dF(u)
\]

A fundamental characteristic of the DP model is that, given a sample \( X_n = (x_1, \ldots, x_n) \) drawn independently from (uncertain) distribution \( F \), the posterior is the DP

\[
F \mid X_n \sim \text{DP}(\alpha + n, F_n), \quad F_n = (\alpha F_0 + \sum_{i=1}^{n} \delta_{x_i}) / (\alpha + n).
\]

\[
\mathbb{E}[f \mid X_n] = a_n \int k(x, u) \, w(u) \, dF_0(u) + n^{-1} (1 - a_n) \sum_{i=1}^{n} w(x_i) \, k(x, x_i)
\]

where \( a_n = \alpha / (\alpha + n) \). Taking the formal limit of \( \alpha \to 0 \) to represent a non-informative prior leads to the finite-dimensional Bayesian Representer form

\[
\hat{f}_n(x) = \sum_{i=1}^{n} w_i \, k(x, x_i),
\]
The model

• Likelihood:

\[ y_i = w_0 + f(x_i) + \varepsilon_i = w_0 + \sum_{j=1}^{n} w_j k(x_i, x_j) + \varepsilon_i, \quad (i = 1, \ldots, n) \]

\[ Y \sim N(w_0 t + Kw, \sigma^2 I) \]

• Prior:

\[ (w \mid K, T) \sim N \left(0, F \Delta^{-1} T \Delta^{-1} F' \right) \]

\[ K = F \Delta F' \]

\[ \Delta = \text{diag}(\lambda_1^2, \ldots, \lambda_n^2) \]

\[ w = F \Delta^{-1} \beta \]

\[ \beta \sim N(0, T) \]

\[ T = \text{diag}(\tau_1, \ldots, \tau_n) \]

\[ \tau_i \sim \text{InvGa}(a_\tau/2, b_\tau/2), \quad (i = 1, \ldots, n) \]
Inference – Gibbs sampler

to simulate the posterior $p(w_0, w, \sigma^2 \mid \text{data})$

1. Update $w_0$: $w_0$ is drawn from the normal posterior with mean $n^{-1}i'(Y - F\beta)$ and variance $\sigma^2/n$.

2. Update $w$: Simply via $\beta$, generate $\beta \sim N(b, V)$ where $V = \text{diag}(V_1, \ldots, V_m)$ with $V_i = \sigma^2\tau_i/((\tau_i + \sigma^2)$, and $b = V F'(Y - w_0)/\sigma^2$; then set $w = F\Delta^{-1}\beta$.

3. Update $T$: For $j = 1, \ldots, m$, $\tau_j^{-1} \sim \text{Ga}((a_\tau + 1)/2, (b_\tau + \beta_j^2)/2)$.

4. Update $\sigma^2$: $\sigma^{-2} \sim \text{Ga}(n/2, s/2)$ with $s = e'e$ where $e = Y - w_0 - F\beta$. 
Results - The MNIST Data Set

Figure 4: The MNIST data. Plot of the 45 classification errors for the kernel model with variable selection (solid line with circles) and without variable selection (dashed line with stars).
Figure 5: The MNIST data. Upper panel: plot of relevant variables (the 1st, 2nd, 4th and 5th variables). Lower panel: plot of irrelevant variables (the 3rd, 6th, 10th and 11th variables).
Figure 6: The MNIST data. Plot of projections onto sets of two relevant variables, where circle represents “3” and square represents “5”. The two classes show some separation in the relevant variables.

Figure 7: The MNIST data. Plot of projections onto sets of two irrelevant variables, where circle represents “3” and square represents “5”. The two classes are mixed in the irrelevant variables.